The Winemakers Dilemma – Part 1

The process of making wine is well known and has been an active industry for many centuries. What is not as well known is that this process is fraught with risk. Between extreme/established competition, weather dependence, equipment maintenance, and regulation most new wineries fail[i]. While the wine industry is growing and the number of wineries is growing,[ii] there is a move towards consolidation with a small number of large players dominating much of the market. However, with size comes a quantity-quality trade-off that keeps much of the niche markets fertile ground (pun intended) for smaller wineries.

Alejandro Martinez is a new vineyard owner. He purchased a small pre-existing vineyard and winery in the Columbia River basin in eastern Oregon currently producing 10,000 cases of wine annually. The winery is called Diplomatico Vineyards (a nod to his family's Venezuelan rum making roots) and specializes in various German style Rieslings. Specifically, there is a spectrum of Riesling types that are created using the white grapes grown on the property[iii]. In general, the sweeter the more expensive. The following is a list of wine types, and the typical quantity of cases produced each year at Diplomatico.

- Trocken (dry) - 1000
- Kabinett (dry to off-dry) - 1000
- Spätlese (sweet) - 5000
- Auslese (sweeter) - 1000
- Beerenauslese (very sweet) - 1000
- Trockenbeerenauslese (super sweet) – only produced by exception

In general, the sweeter the more expensive. In fact, the some of the most expensive wines in the world are of the Trockenbeerenauslese variety (more on this later)[iv]. However, to get these sweeter wines, the winemaker must harvest the grapes later when at their ripest and the sugar content is highest and acidity levels are lower. The dryer wine (Trocken and Kainett) has grapes harvested with 20% sugar, where the sweeter wines (Spätlese, Auslese, and Beerenauslese) have closer to 25% sugar. However, this late harvest can be risky as waiting too long can spoil entire crop.

As a former technology executive, Alejandro hopes to overcome obstacles faced by upstart wineries with a data-driven focus. However, he knows this is easier said than done. Most wineries rely on expert knowledge from a small and expensive collection of fermenters, harvesters, and sommeliers. With a modest budget he hopes to replace some of the high-cost experts with business intelligence, predictive analytics, and machine learning. However, he must prioritize the right data projects and infrastructure to make the best use of his limited resources.

The Data Eco-System

You have been hired by Alejandro as the data science lead. Currently you have a 3-person team, including yourself. A data engineer and full stack software developer are also on staff. Your first task is to build a data eco-system that collects relevant data so that you can be ready to support decisions when the time comes.

Task 1: Identify the following data management / MLOps components that would be of most benefit to the vineyard:

- Data Management System
    - Sensors (data sources)
    - Data Pipelines
    - Storage
- Data Analysis/Science
- Product Delivery

Specify what data you would collect to help the winemaker create informed decisions. Graphically, design (using PowerPoint or similar) a basic data eco-system, complete with recommended technologies, that can be used to collect, analyze, and model with this data.

Task 2: One of the major determining factors of wine quality is the weather. Collecting specific weather data will be a critical source of value. Alejandro has paid for access to the following API:

- https://openweathermap.org/api

Create a data pipeline that collects current weather data for Hood River, Oregon. This pipeline should be a single python script that collects the data performs an ETL process and saves it into a SQLite database (provided). In your script include comments on your approach and considerations with scaling. Specifically, answer the following questions (again as comments in your script).

- What, if anything, would you do differently if you knew this script was going into a production environment?
- What are some other options for persisting this data? Is a structured table the best selection for this application? Why or why not?
- How did you make your approach 'readable' for other data engineers? Are there some common design patterns you used or wanted to use that would be helpful?
- How might you standardize this data pipeline development process across a diverse range of data types?

---

[i] https://www.northbaybusinessjournal.com/article/industry-news/5-big-risks-to-wine-businesses/
[ii] https://wineindustryadvisor.com/2022/01/27/wine-distribution-solutions
[iii] http://socialvignerons.com/2018/05/17/types-aromas-stories-behind-riesling-wines/#riesling-types
[iv] http://socialvignerons.com/2017/10/06/top-50-most-expensive-wines-in-the-world/